**A formal proposal for a MSc project that will be submitted in partial fulfilment of a University of Greenwich Master’s Degree**

**Identifying insight and pattern from Britain Outbound and Inbound tourism and produce a user interactive information interface dashboard using Power Bi**

**Name: AYOMIKUN HAKEEM LAWAL**

**Student ID: 001186392**

**Course of Study: BIG DATA & BUSINESS INTELLIGENCE**

**Supervisor: Dilan ThelikadaPalliya Guruge**

**Topic Area:** Leisure & Tourism – Big Data & Business Intelligence

**Keywords associated with the project:** Dataset, Data Analysis, Data Science, Travel.

**MSc Modules studied that contribute towards this project:** Data Visualisation, Data Warehousing & Business Intelligence, Machine Learning.

Table of Contents

[Abstract 3](#_Toc123957455)

[1. Introduction 4](#_Toc123957456)

[1.1 Problem Statement 5](#_Toc123957457)

[1.2 Aim and Objectives 5](#_Toc123957458)

[2. Background 8](#_Toc123957459)

[2.1 Literature Review 8](#_Toc123957460)

[2.2 Review of Existing project 9](#_Toc123957461)

[3.1 Methodology 10](#_Toc123957462)

[3.2 Review of tools and techniques 13](#_Toc123957463)

[4.1 Implementation 14](#_Toc123957464)

[4.2 Analysis Discussion 16](#_Toc123957465)

[4.3 Time series Forecasting Discussion 22](#_Toc123957466)

[4.4 Machine Learning Model Fitting for Holiday Package 27](#_Toc123957467)

[4.4 Power Bi Dashboard Design 34](#_Toc123957468)

[5.1 Conclusions 38](#_Toc123957469)

[References 40](#_Toc123957470)

[Appendix 41](#_Toc123957471)

# Abstract

This study addresses the need to present an excellent travelling information from passengers travelling into the UK who reside both in the UK and the overseas. It has been observed that there is a less or no platform where an analytical insight or information can be retrieved as regards travelling into the UK. The problem of insufficient information can make the country vulnerable to certain risk, such as insecurities. The proposed study will promote economic and security growth, environmental issues will be affected positively and transportation crisis would be reduced.

The chosen method of the study is data analysis and visualization which was done using python programming language. The foundation knowledge for this study is the dataset used. The dataset consists of the quarterly information of passengers travelling in and out of the UK. It includes their age, sex, purpose of trip etc. This dataset is primarily collected from the International Passenger Survey (IPS). The International Passenger Survey (IPS) is a continuous survey carried out by the Office for National Statistics (ONS) that collect information from travelling passengers in the UK. This dataset would be analysed and several insights was derived from it. The dataset and results retrieved was further used in creating a dashboard using Power Bi.

The key findings in this study are the insight derived from the dataset and how useful the result can be applied to solve real life issues. The dashboard created can help retrieved information at a glance and also determine some travelling factors and facts about the passengers.

***Keyword:*** *Time series, Arima, Power Bi, Forecasting*

# 1. Introduction

There are three modes of transportation in the UK: land, water, and air. Every form of transportation that takes place on the ground is considered land transportation, including the use of an automobile, truck, train, etc. On the other hand, water transportation refers to travel on bodies of water and entails the use of boats, ships, etc. The final method is air transportation, which involves moving people or goods by plane or helicopter. Air and water/sea transportation in the UK can encompass both domestic and international travel, whereas land transportation mostly concerns domestic travel. In this study, we'll concentrate on both water (the sea) and air transportation for travellers departing the nation (UK).

A dataset is provided for this study to perform an analysis. The dataset consists of a series of data files that enable users to conduct analysis of overseas travel and tourism estimates within certain subgroups. The subgroups can be related to passengers of the same country, sex, age group etc. The data files are derived from the International Passenger Survey (IPS). The International Passenger Survey (IPS) is a continuous survey carried out by the Office for National Statistics (ONS). It covers all major air, sea, and tunnel ports, providing detailed information on the numbers and types of visits made by people travelling to and from the UK. Data is published regularly by ONS on a monthly, quarterly, and annual basis. Anonymous face-to-face interviews are undertaken with a random sample of passengers as they enter or leave the UK. Approximately 95 per cent of passengers entering and leaving the UK have a chance of being sampled on the survey.

It consists of information of international passengers such as the total night spends while on the visit, the total amount of money spends, the purpose of visit, the mode of transportation into the country, the passengers age, the length of stay during visit, the place of stay during visit etc. All this information is collected individually from the passengers inform of a raw data and few information can be deduced from the dataset without being analysed, this has led to the development of this project.

The main aim of this work is to analyse the dataset and derive several insights from the dataset to produce more meaningful information. Before proceeding on extracting insights, the dataset would undergo data preparation stages, the process involves checking the dataset to see if it’s fit for the experiment, we check if there are any null values in the dataset or empty data values. At the end of the process, the dataset must be ready for the experiment. A graphically visualization of the analysis would be done using Power Bi

## 1.1 Problem Statement

In our increasingly data-driven world, it’s more important than ever to have accessible ways to view and understand data. According to the research and survey conducted, there is no available statistical dashboard where information regarding travelling in and out of the UK can be collected.

Existing papers focuses on transportation challenges or development of policies that would govern fair transportation. Often times, transportation agencies focus more on any condition or activities such as weather, theft, disaster that can affect the transportation itself and situation that can cause delay or termination of the transportation without primary collecting information from the passenger or providing insight to ensure safe and secure transportation.

That’s where data visualization comes in handy. With the goal of making data more accessible and understandable, data visualization, in the form of dashboards is being used by many businesses to analyze and share information. In this study, the dataset collected will contain information about passenger who are travelling in and out of the UK, insight would be derived from the dataset and a data dashboard would be developed where more information can be represented.

## 1.2 Aim and Objectives

To achieve the aim of this study, the following objectives would be considered.

* 1. To obtain a passenger information dataset. The dataset was downloaded from <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/travelpac> . The dataset consist of the quarterly information of passengers travelling in and out of the UK. It includes their age, sex, purpose of trip etc. This dataset is primarily collected from the International Passenger Survey (IPS). The International Passenger Survey (IPS) is a continuous survey carried out by the Office for National Statistics (ONS) that collect information from travelling passengers in the UK.
  2. To conduct a requirement gathering for the successful implementation of this work. This process involves reviewing of related paper and conducting a research as regards the best tools to be used to carry out the data analysis, the visualization and the web dashboard. The functional and non-functional requirement would be identified in other to know the function of the proposed system and its efficiency.
  3. To perform data cleaning on the dataset downloaded. Data cleaning involves checking the dataset if it is fit for the experiment. The processes involve checking for null and empty values, it also involves checking if the data is consistent and free from any kind of errors. This cleaning process would be carried out using programming techniques like python and Microsoft excel.
  4. To analyse the dataset. This involves using programming and statistical tool to extract insight from the dataset. Some of the insight that would be generated from the dataset are;

**Insight about passengers and their country.** This insight can help identify the country with the highest and the lowest number of passengers. Also, the dataset can be grouped by the country of resident to perform inter country analysis. This can work with all other metrics like the age or gender, to know which of the gender travel the most. Also, the average age of passenger who are travelling.

**Insight relating to duration of night spent.** This would help to identify the average number of nights spent by the passenger. This can be grouped by country as well for inter country analysis also can be grouped and analysed by the purpose of visits. Information such as the average number of nights spend by passengers who came for holidays, business, study etc. can be identified. This can help predict expenditures for future passengers with the same purpose of visits. This information can be useful to hotel managers or real estate agencies in the resident to plan themselves and provide space for incoming passengers which can also help in the promotion of business.

**Insight from mode of travel.** From the metric of mode of travel, we can derive insight like the most used mode of travelling by passengers. This can also help the government in identifying the means of transportation which is mostly used in case of emergencies or identifying which means of transportation to focus more on or invest on.

**Insight from package**. This identifies whether the passenger travelling is based on a tour package or travelling independently. The dataset can be classified based on this metric to identify which passenger fall in the independent and non-independent category. This metric can work with the duration and the spend metric to derive more insight like the number of night and the expenditure of independent and non-independent passengers and whether the purpose of visit affect independent or non-independent passengers.

**Insight from country.** The country can be used to derive insight of passengers relating to a particular country, like the most common mode of transportation in a country, countries most traveling purpose and their expenditures. This metric can provide more economic insight about country, relating to the purpose of visits.

This insight can help researcher and travelling agencies to be informed about information that cannot be seen on a regular customer data, also for future research purpose, this can serve as a steppingstone for future research works about traveling in and out of the UK. This project would be useful in situation where more insight or information is needed about overseas visitors to the UK for tourism policy and provide data to feed into estimates of international migration. More insights can be extracted by the combination of several columns and predictions can be done on the data as well.

* 1. After several insight has been extracted from the dataset. Data visualization would be implemented using Power BI. Data visualization is the process of graphically representing the data to showcase the results of the data analysis as well as the dataset itself. The visualization would contain charts such as bar chart, pie chart and every other chart depending on the results of the insight discovered.
  2. A dashboard would be created using the visualization done with power bi. This dashboard would present an interface where all information and insight can be seen at a glance. This dashboard would contain different chart such as bar chart, stacked bar chart, box plot, graphs.

# 2. Background

## 2.1 Literature Review

Several research papers with similar aim and objectives would be discussed in this section.

Analysis from the National Travel Survey (NTS) by (Drren Stilwell, 2018)is a household survey of individual travel by residents of England within Great Britain. The data used was gathered through interviews and a one-week travel diary. The NTS is a part of a survey that started in 1988 which allows for the investigation and the discovering of patterns and trends in England. The data collected can be used for a number of important purposes, such as describing patterns, which includes how different groups of people travel, tracking travelling trends, like sustainable modes, assessing the potential equality effects of transportation policies on various groups, and helping to evaluate the effects of policies. In the survey, it was observed that there have been substantial changes in travel behaviour since 1975. Between 1975 and 1990, the annual average number of travels per person rose, and since 1995, it has been declining. In comparison to the Greater London Built-up Area, distance travelled per person increased by 80% and by three times in the smallest towns and rural areas between 2011 and 2014. The amount and type of modes utilized in multi-mode travels vary for London, urban and rural locations, and depending on the purpose of the trip, even though a relatively small percentage of trips do so. The Strategic Road Network (SRN), local roads, rail, and buses all generally received equal levels of satisfaction. However, satisfaction is higher for walking provision and lower for cycling provision (both at 27%) (73%).

(Halden, 2019) UK Travel Time, Accessibility and Connectivity Statistics. This article contributes to the ongoing government assessment of statistics by discussing the evolution of travel time, accessibility, and connectivity data.The aim of this study is to make transport system very accessible. The articles also describe the opportunity available to transport users and non-users, it also provide a solution to the factor affecting travel cost and distance.Finally, the importance of collecting transport data was justified and several usage illustration were identified.

(Harms et al., 2018)in an article titled, using time-use data to analyse travel behaviour: Findings from the UK used a survey called Time Use Survey (TUS) which aimed at collecting detailed information about individual daily activities involving travelling to analyze travel behavior. This study also examines and discuss the capability of the TUS data and how useful it is to research papers that involves transport. The study compared TUS and activity travel diaries to achieve the aimed analysis. In conclusion, it show that TUS can successfully supplement (national) travel surveys in their transition to a more effective activity-based travel models by providing rich data (specific travel purpose, contextual information, and enjoyment levels).

In the study on the quality of the travel experience at city destinations, (Yuniawati&Ridwanudin, 2015)sought to develop a method for evaluating the tourist experience in three phases: the pre-travel phase, which was assessed using the information search construct, planning, and decision-making; the experiential phase, which was assessed using the product attributes, interaction, and involvement constructs; and the recollection phase, which was assessed using the meaning, satisfaction, and satiety constructs. Factor analysis and data analysis techniques are used in the descriptive and verification phases of the research methodology. As many as 200 tourists that visit the city of Bandung were surveyed and interviewed to gather data. The findings demonstrated that the suggested model is accurate and valid for assessing the travel experience.

(Alit Suthanaya, 2018)did an analysis of travel pattern and the need to develop sustainable transportation infrastructure in Sarbagita metropolitan area. With the maximum population and activity of the city, it is difficult to understand travel pattern.

In order to create a more sustainable transportation infrastructure, this study looks at current travel patterns as well as predictions for the future. The origin-destination of persons and things, as well as statistical population data, are the data used. Using Visum software and four phases of modeling, the prediction approach was validated using average daily traffic (ADT) data. According to the study's findings, traffic builds up on most arterial and collector roads, especially those around Denpasar, with a saturation level above 1.

## 2.2 Review of Existing project

Similar platforms that helps in providing information about international transportation such as; International Transport Forum (ITF) <https://www.itf-oecd.org/about-itf>. Travel State Gov <https://travel.state.gov/content/travel/en/about-us.html> and Travel Analytic Center<https://support.google.com/travel/>. The ITF which is a discussion platform that uses feedback from the forum to produce the information its generate on it platform. The system also analyses trend but the information used are extracted from the discussion. The problem associated with this is the mode of data collection and the platform does not have a statistical visualization of expected trends.

The travel state gov is a travel policy agency whose priority is to provide safety and security service and to protect lives of passengers in the U.S. They conduct the practices through routine and emergency service to American embassies. This platform only focuses on the security of the passengers. While the travel analytic centre is a platform designed by google to help travellers organize their trips, the flight to take, travel advice and also book an hotel respective destination area. This platform is designed to provide services to the passenger. The information it provides are limited as it concerns only rendering services to passengers. This problem necessitates the development of this study to show how insight and pattern can be extracted from passengers’ information.

The difference between existing system and this proposed system is to develop a system that will provide information which can guide against insecurity of the country involved in the transportation as well as the passengers using the transportation. The system will promote economic and security growth, environmental issues will be balances and transportation crisis would be reduced. This process would be carried out with the data and information collected from travelling passengers using IPD Travelpac dataset, IPS (International Passenger Survey) aims at conducting surveys of international passengers. This survey helps collect basically raw information from international passengers. The dataset is organized monthly or quarterly. Analysis of the passenger information to identify insight and pattern will be done and a well presentable dashboard for data visualization would be developed. With the new development incident or outbreak can be controlled if there are proper information from the people traveling into the country.

# 3.1 Methodology

In this study, several methods or steps would be followed to achieve the aim of this study. The methodology covers the steps to be taken and to derive insights from the dataset and also creating the dashboard.

Below are the methods to be carried out.

1. **Data collection:** This section involves the dataset to be used. The data files are derived from the International Passenger Survey (IPS). The International Passenger Survey (IPS) is a continuous survey carried out by the Office for National Statistics (ONS). It covers all major air, sea, and tunnel ports, providing detailed information on the numbers and types of visits made by people travelling to and from the UK.

Anonymous face-to-face interviews are undertaken with a random sample of passengers as they enter or leave the UK. Approximately 95 per cent of passengers entering and leaving the UK have a chance of being sampled on the survey. It consists of information of international passengers such as the total night spends while on the visit, the total amount of money spends, the purpose of visit, the mode of transportation into the country, the passengers age, the length of stay during visit, the place of stay during visit etc. All this information is collected individually from the passengers inform of a raw data and few information can be deduced from the dataset without being analysed, this has led to the development of this project.

The dataset consists of a series of data files that enable users to conduct analysis of overseas travel and tourism estimates within certain subgroups. The subgroups can be related to passengers of the same country, sex, age group etc.

The dataset was downloaded from this source <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/travelpac>

1. **Data Cleaning**: Data cleaning is the process of removing data from a dataset that does not belong there. It aids in the correction or elimination of inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset. Inaccurate data makes outcomes and algorithms untrustworthy. Because the procedures differ from dataset to dataset, there is no one definitive way to specify the precise phases in the data cleaning process. This cleaning process would be applied on the dataset downloaded to ensure an excellent results and python programming language would be used.

In the Python programming language, the pandas library provides useful tools for data cleaning. One way to clean data using pandas is to use the isnull() function, which returns a Boolean dataframe indicating which values are missing. We used the sum() function to count the number of missing values in each column. We also used the dropna() function to drop rows or columns with missing data. Another technique for data cleaning, outlier detection and removal. These techniques can helped to handle missing or invalid data and prepare our dataset for analysis. Previewing of the dataset would be done using Micro-soft excel.

1. **Data Analysis:** This is the process of generating insights from the datasets. This process would be done using python programming language and Jupyter note book which is the IDE (Integrated Development Environment) to write python codes.

Transportation related insights would be derived, such as information about the most popular mode of transportation for passengers from the metric of mode of travel. This can assist the government in determining the modes of transportation that are most frequently used, as well as the modes of transportation it should concentrate its efforts and resources on.

Insight from package: This identifies whether the passenger travelling is based on a tour package or travelling independently. The dataset can be classified based on this metric to identify which passenger fall in the independent and non-independent category. This metric can work with the duration and the spend metric to derive more insight like the number of night and the expenditure of independent and non-independent passengers and whether the purpose of visit affect independent or non-independent passengers.

Insight from country**:** The country can be used to derive insight of passengers relating to a particular country, like the most common mode of transportation in a country, countries most traveling purpose and their expenditures. This metric can provide more economic insight about country, relating to the purpose of visits**.** This would help to identify the average number of nights spent by the passenger. This can be grouped by country as well for inter country analysis also can be grouped and analysed by the purpose of visits. Information such as the average number of nights spend by passengers who came for holidays, business, study etc. can be identified. This can help predict expenditures for future passengers with the same purpose of visits. This information can be useful to hotel managers or real estate agencies in the resident to plan themselves and provide space for incoming passengers which can also help in the promotion of business.

Insight from mode of travel**:** From the metric of mode of travel, we can derive insight like the most used mode of travelling by passengers. This can also help the government in identifying the means of transportation which is mostly used in case of emergencies or identifying which means of transportation to focus more on or invest on.

Insight from package. This identifies whether the passenger travelling is based on a tour package or travelling independently. The dataset can be classified based on this metric to identify which passenger fall in the independent and non-independent category. This metric can work with the duration and the spend metric to derive more insight like the number of night and the expenditure of independent and non-independent passengers and whether the purpose of visit affect independent or non-independent passengers.

Package is a categorical variable, a machine learning model would be fit to predict the type of package.

Insight from country**:** The country can be used to derive insight of passengers relating to a particular country, like the most common mode of transportation in a country, countries most travelling purpose and their expenditures. This metric can provide more economic insight about country, relating to the purpose of visits.

The main techniques to be followed to derive insight from the data are:

Segmentation. Segmentation is the act of breaking the dataset into smaller junks or categories, by doing this, mining the data would be easier. The insights derived can assist researchers and travel firms in learning about information that cannot be seen in typical customer data.

1. **Data Visualization:** Data visualization is the practice of putting information into a visual context, like a map or graph to make data easier for the human brain to grasp and draw conclusions from. Data visualization's major objective is to make it simpler to spot patterns, trends, and outliers in big data sets.

The cleaned dataset and the result from the analysis would be used to create the visualization dashboard using Power BI software.

## 3.2 Review of tools and techniques

To achieve the aim of this study, the following tool and techniques would be used. The tool is divided into two, which are the software and the programming techniques to be used.

For the software, we have;

1. **Anaconda Software**: Anaconda is a software distribution of the Python and R programming languages to perform scientific computing, which aims to simplify package management and deployment. It will guide in the installation of all the library necessary to achieve the aim of this project.
2. **Jupyter Notebook**: This is a text editor that is available in Anaconda software. It is an IDE designed to write python codes. The experiment would be coded on Jupyter Notebook.
3. **Programming techniques:** The techniques include the programming language, libraries and the algorithms to be used. The programming language to be use is python programming language. Python is one of the most accessible programming languages available because it has simplified syntax and not complicated, which gives more emphasis on natural language (Kelly, 2019). This has made the analysis easy to achieve the aim of this study. Also for the analysis, python has a built-in libraries, which would help in reducing the complexity of certain task. There are several popular libraries in Python for data science tasks, including:

*NumPy*: A library for working with large, multi-dimensional arrays and matrices of numerical data.

*Pandas*: A library for working with data frames, which are tables of data with rows and columns. Pandas provides functions for reading and writing data, filtering and sorting data, and handling missing data.

*Matplotlib*: A library for creating visualizations of data, such as line plots, bar plots, and scatter plots.

*Scikit-learn*: A library for machine learning tasks, including classification, regression, clustering, and dimensionality reduction.

*Seaborn*: A library for creating statistical plots, such as box plots and violin plots.

These libraries are widely used in the data science community and would be used in this report.

# 4.1 Implementation

The implementation of this study, covers deriving an insight from the dataset used and creating a dashboard using power bi. The process starts by launching anaconda, which has Jupyter notebook installed in it. Jupyter note book is a text editor design to write python codes.

To begin the experiments, it starts by importing the library needed. The first library imported is pandas, pandas is used for data manipulation. The pandas were used in importing the dataset. Numpy was used for arithmetic computation. The seaborn and matplotlib works together to plot the charts needed. After the library has been imported, the next step is to use pandas to import the dataset and assign the dataset to a data frame. The dataframe is a variable name assigned to the dataset. The dataframe can be previewed and cleaned. Data cleaning involves checking if there are any bias values in the dataset. This can be checked using a python command and can be seen in the appendix section.

**4.2 Data Cleaning**

From the raw dataset. It shows that in column sex, visits, nights and sample contains null values which cannot be empty cells. Because we have a large number of dataset samples, deleting the row with any null values would not have much impact on the result of the analysis. The entire dataset has 425,024 samples with 14 columns before data cleaning. After removing the null values the number of samples decreased to 410,573 samples. A thorough checking of the values of each columns showed that columns like sex, age and duration of visits contains some unwanted inputs. Sex contains “Don’t know and ‘ ‘ ,an empty quote” , age contains “D/K” as values while duration of visits contains “Stay not Known”.

After the unwanted values were removed the number of samples decreased to 404,775.

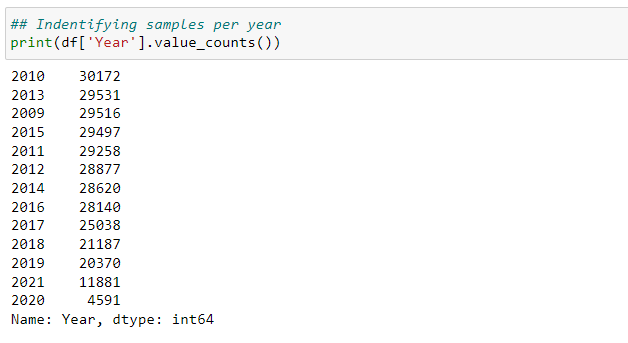
The cleaned dataset was exported to a new data sample and then re-imported and begin the process of removing outliers.

**4.3 Removing outliers**

Outliers, or extreme values that are significantly different from the majority of data points, can have a significant impact on the analysis and interpretation of a dataset. They can skew statistical measures such as the mean and standard deviation, and can also affect the results of statistical tests and models. Outliers can also lead to errors in data visualization, such as distorting the scales on a graph. It is important to identify and handle outliers appropriately in order to accurately understand and interpret the data. A box plot of some columns showed that there are outliers in them. The outliers were removed and this reduced the dataset to 316,678 samples.

**4.4 Identifying samples per year**

From the entire dataset. 2010 has the highest number of data samples, while 2020 has the least as shown in fig 1 below



*Figure 1:Dataset samples per year*

# 4.2 Analysis Discussion

The insight derived is as a response to some research questions, such as;

1. What are the insights related to the travelers as regards their country, night spent, mode of travel and package of travel?
2. What necessitates travelling?
3. Does age groups affects travelling?

In other to make the analysis easier , and for proper insight generation. I concluded to segment the data. Segmentation helps me to break down all the data into smaller chunks to simplify the complexity of the analysis. The first segmentation done is breaking the dataset into two categories, using the ukos (where\_contact\_lives) variable which identifies where contacts lives. With this, I now have dataset of travelers who lived in the UK and the travelers who lived in the overseas.

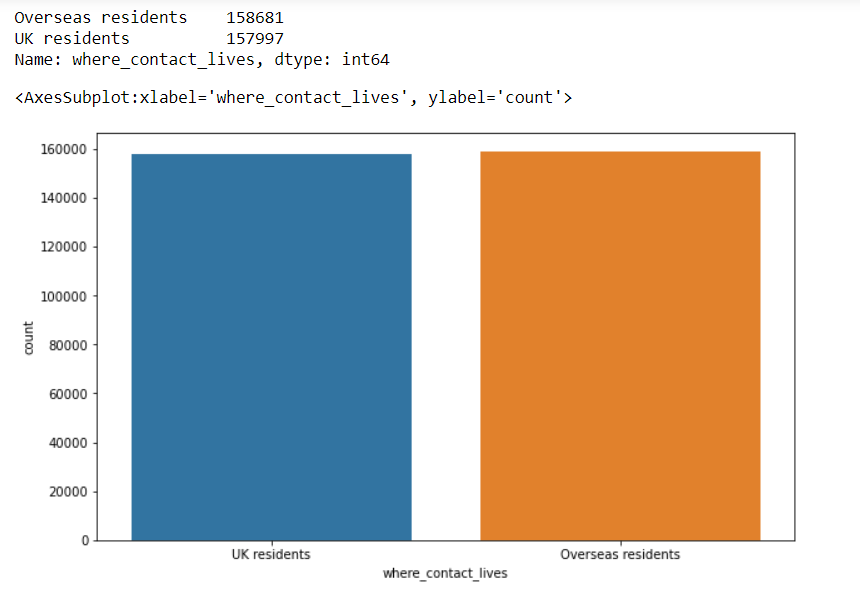


Figure 2: Where contact lives

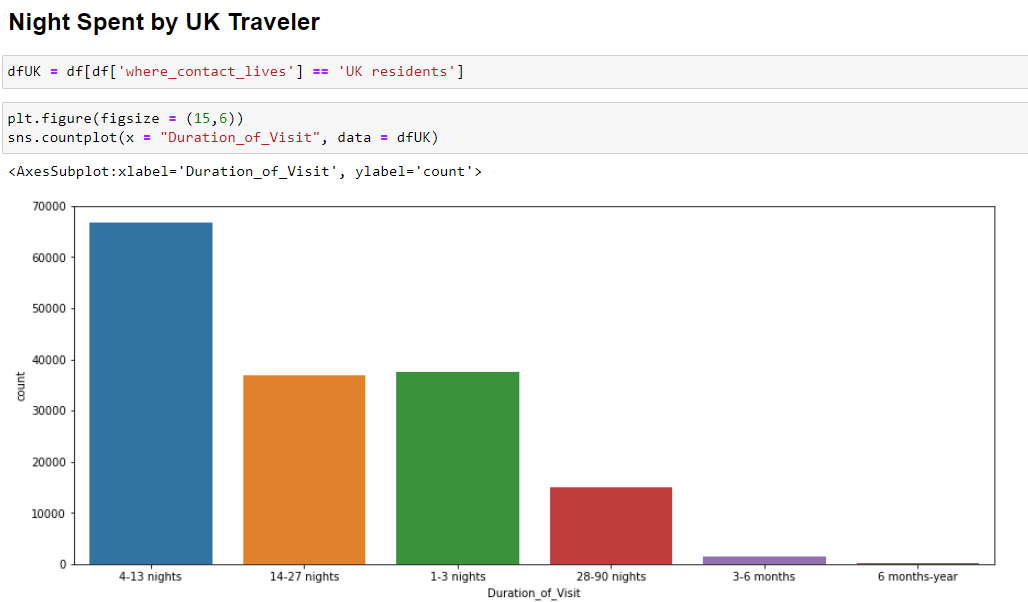
Fig 2 shows that from the overall dataset we have more oversea travelers than the UK travelers. Going further with the classification to identify what necessitates travelling, I considered identifying how many nights spend between both travelers. This process involves comparing the result from the UK and the Overseas travelers to identify the similarities and differences and detect the insights. 

Figure 3: Night spent by UK Traveler

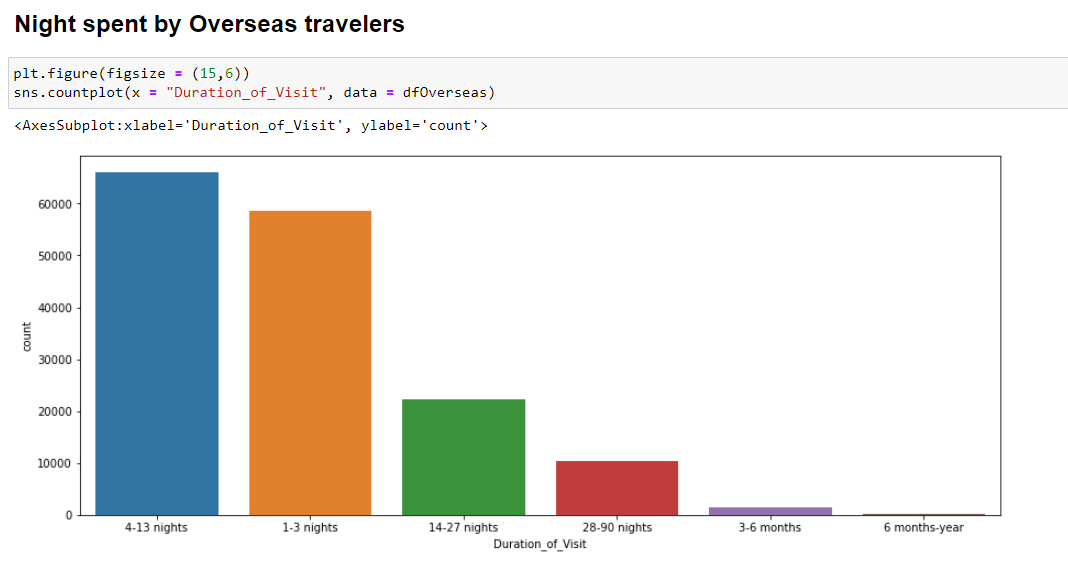


Figure 4: Night spent by Overseas travelers

The bar chart above shows of all the night available, travelers who lives in the UK and overseas travelers spent 4 to 13 nights more ,the number of UK travelers who spent 14-27 nights and 1-3 nights hovers around the same mean, while the margin between oversea travelers who 1-3 nights more and 14-27 nights is so wide with 1-3 nights more than 14-27 nights. I focused more on these because of its larger population, to identify what necessitate their travelling.

More segmentation was done, where the traveler that spent between 4 to 13 nights was further classified to identify the purpose of travelling.

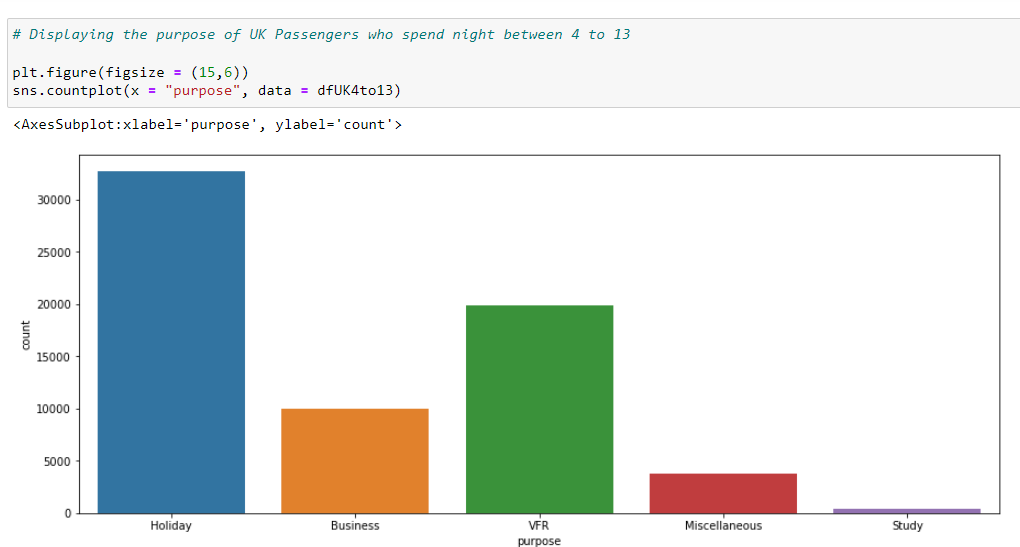


Figure 5: Purpose of passengers who spent 4 – 13 night that stays in the UK

As seen in fig 5, for travelers that stay in the UK, the major reason for travelling is for holiday, the second major reason is visitation.

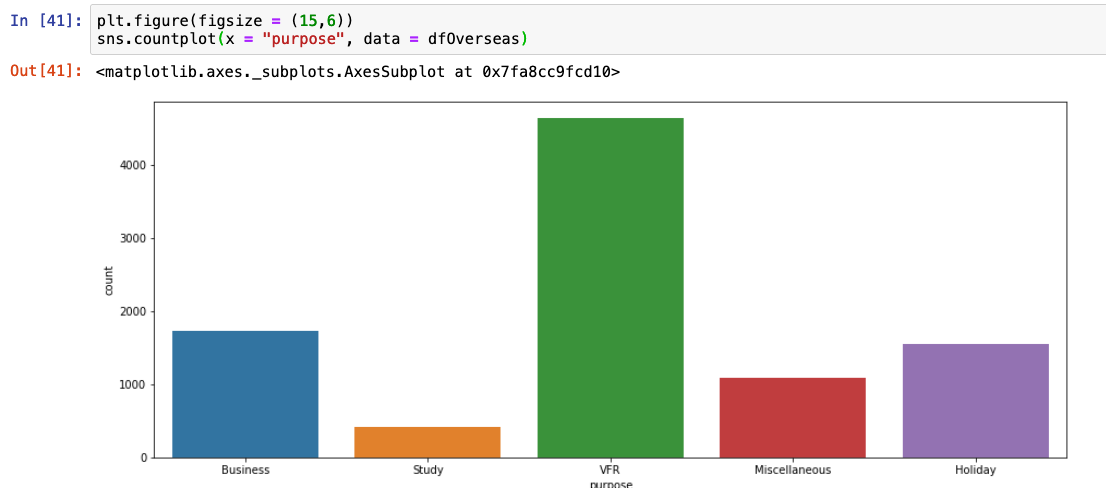


Figure 6: Purpose of passengers who spent 4 – 13 night that stays in the Overseas

As seen in fig 6, it shows that the major purpose of travelling by passengers that spend between 4 to 13 night in the overseas are traveling for visitation purposes while the second major reason is for business. With this information we can conclude that the visitation is one of the major purpose of the travelling either those that lived in the UK or overseas .

Going further, the dataset was grouped by age. This is done to see the kinds of people who travelled for visitation. From the segmentation its shows that traveler between the age 25 – 65 mostly go for visitation. This is very relatable as the age categories covers university graduates who might want to visit home when school are on break or adult who are travelling back home for visitation.

Figure 7: Age groups of passengers who spend 4 to 13 night in the overseas

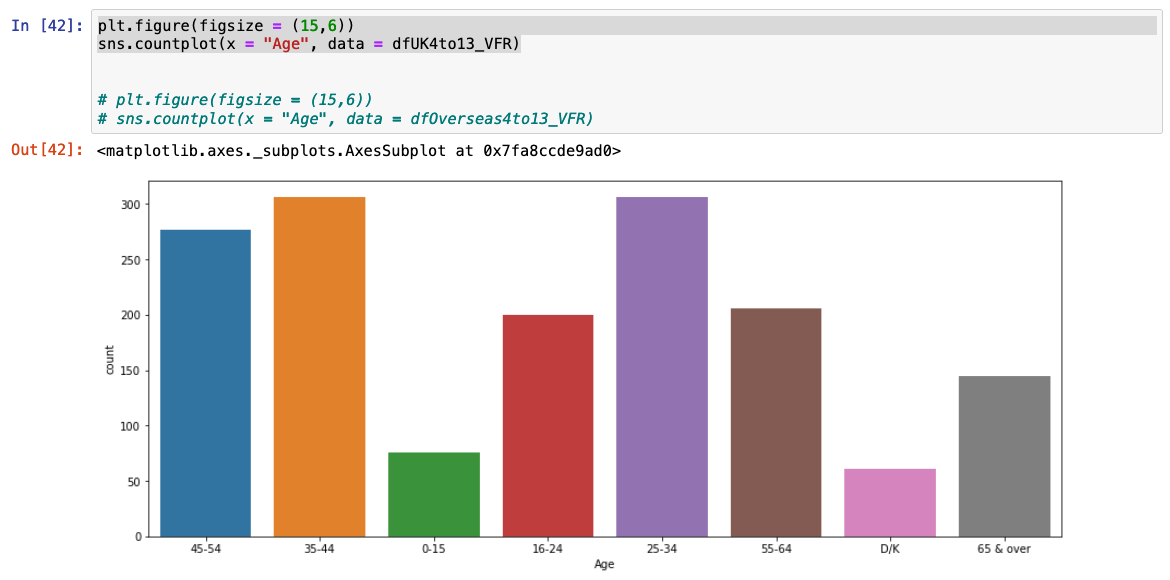


Figure 8: Age groups of travelers who spend 4 to 13 night in the UK.

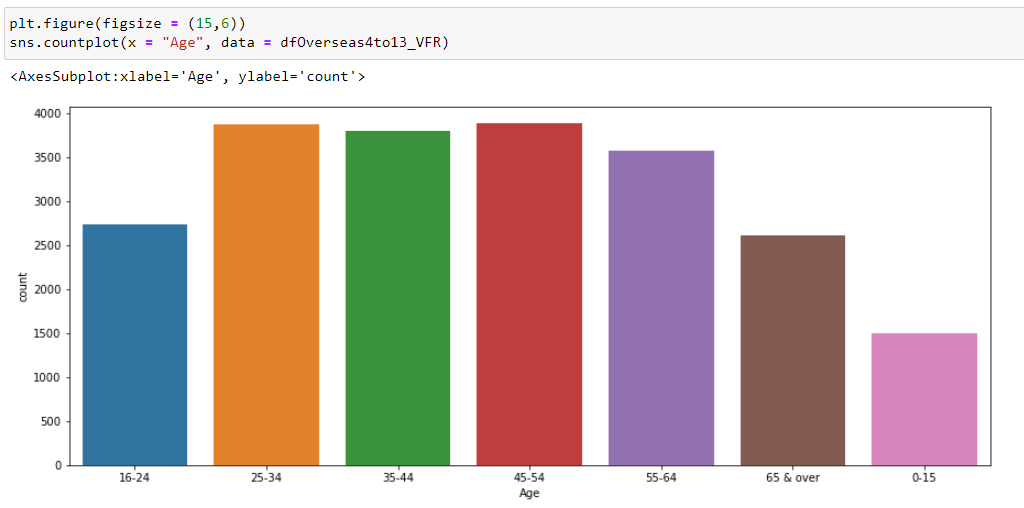


Figure 9: Age groups of travelers who spend 4 to 13 night overseas

From figure 8 and 9 above, there are similarities in the age group of the passengers travelling. From age 25 – 55, we have larger number of passengers in that age group, considering that we have more passengers who travel for visitation and holidays, we can conclude that most passengers between the age of 25 – 55 are travelling for holidays and visitation purpose, more reason passengers are spending 1 to 13 nights during their visits.

Categorizing the dataset into years and identifying the number of visits yearly, as seen in fig below, on a yearly bases, starting from 2009 till 2019, we can see an increase in the number of passengers and a huge drop in the year 2020, this is basically caused by COVID-19 lock-down and in the year 2021, it started increasing gradually.

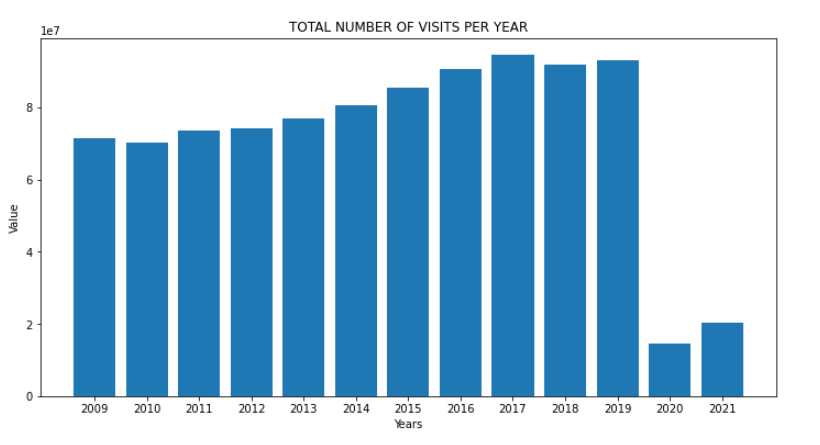


Figure 10: Total number of visits per year

Same analysis is applicable to the number of nights spent as explained above.

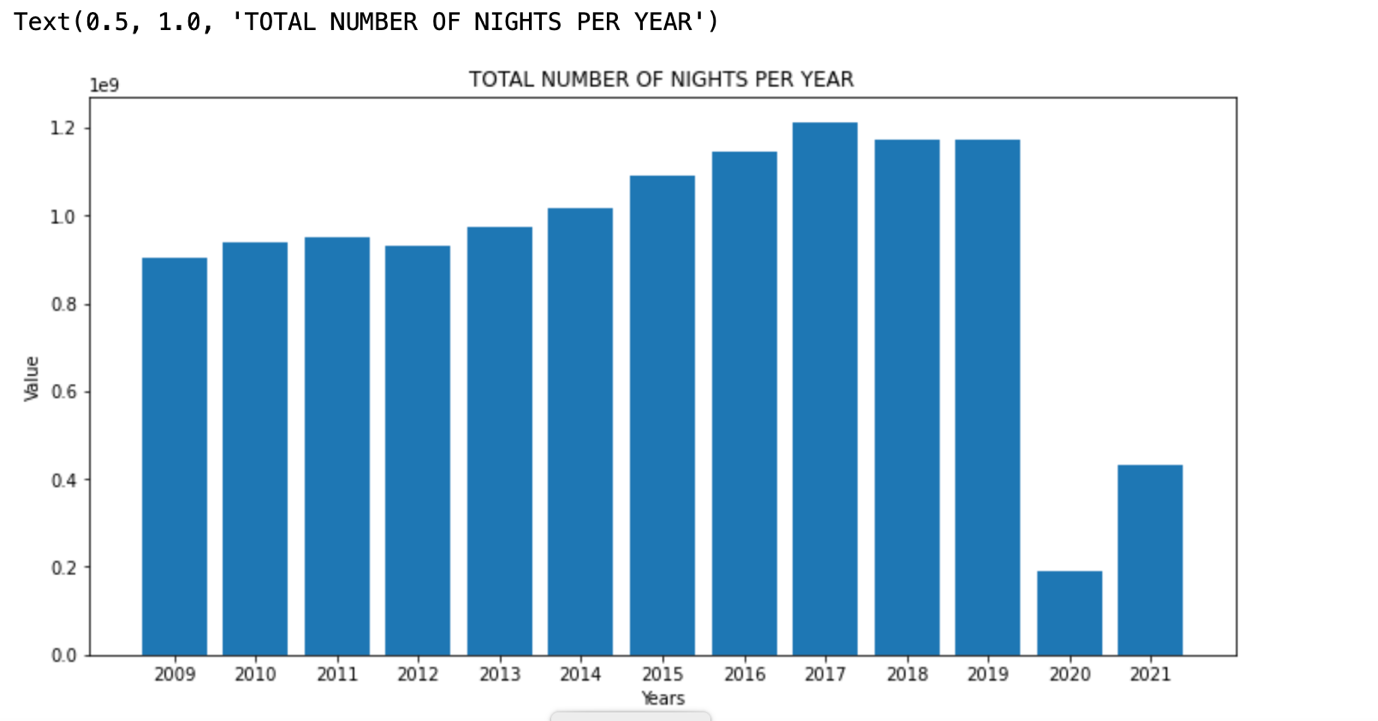
****

Figure 11: Total number of nights per year

From the entire dataset ranging from year 2009 to 2021, It shows that place of residence for overseas residents or of visit for UK residents with the highest number are from France, with 35,021 passengers while Cyprus which is the list has 690 passengers.



Figure 12: Country with the highest number of travelers and Least number of travelers

# 4.3 Time series Forecasting Discussion

A time series is a collection of data points that are ordered. It includes techniques for deriving useful statistics and other aspects of time series data through analysis. This is the foundation for predicting future events by evaluating historical data. An ARIMA model was employed as the forecasting algorithm. A statistical analysis model called an autoregressive integrated moving average, or ARIMA, uses time series data to either better comprehend the data set or forecast future trends. (Mingda, 2018).Time series analysis is being implemented in different field or organization to identify why some specific variables changes over some period of time. It is efficient for stationary dataset i.e., data that the values are prone to changes.Time series analysis is frequently used in the financial, economic, and retail sectors since sales and currencies are continually changing.It is one of the methods used in predictive analytics that shows potential changes in the data, such as cyclical or seasonal activity, which provides a clear understanding of the data factors and aids in improved forecasting.

For this work, time series analysis would be implemented on different variables which are the amount of expenditures of the travelers, the number of nights spent by the travelers and the number of visits of the passengers for the next 5 years. The first experiment done was forecasting the amount of expenditure for the next 5 years.

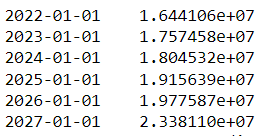


Figure 13:Forecast of Total Amount spent

After the dataset has been imported, it needs to be checked if its stationary or not. Time series analysis relies heavily on the concept of stationarity, which has a major impact on how the data is interpreted and forecasted. Most time series models make the assumption that every point is independent of every other point for forecasting or predicting the future. The dataset of historical instances being stationary is the strongest indicator of this. The statistical characteristics of a system must remain constant across time for data to be considered stationary. This does not imply that the values for each data point must be identical, but rather that the general pattern of the data should not change.

This can be identified using the P-Values of ADF test, which needs to be small as possible, a. higher P-Values shows the dataset is not stationary while a lower P-Values shows the dataset is stationary.

**Augmented Dickey–Fuller Test**

The Augmented Dickey-Fuller Test is used to determine if time-series data is stationary or not. Similar to a t-test, we set a significance level before the test and make conclusions on the hypothesis based on the resulting p-value.

**Null hypothesis and alternative hypothesis for ADF test**

Null Hypothesis: The data is not stationary.

Alternative Hypothesis: The data is stationary.

For the data to be stationary (ie. reject the null hypothesis), the ADF test should have:

p-value <= significance level (0.05).

If the p-value is greater than the significance level then we can say that it is likely that the data is not stationary.

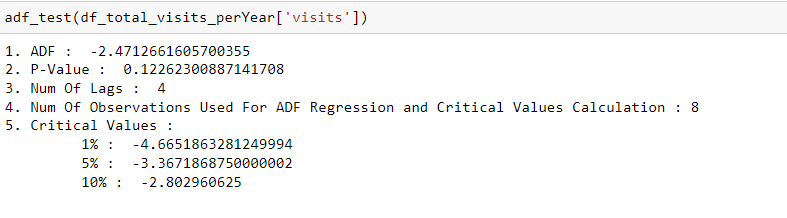


Figure 14: P - Values

From the result above, the P-Value is greater than 0.05. So the null hypothesis accepted, this means that the dataset is not stationary.

A best fit model search was done using auto arima stepwise algorithm, ARIMA(0,2,1) was suggested.

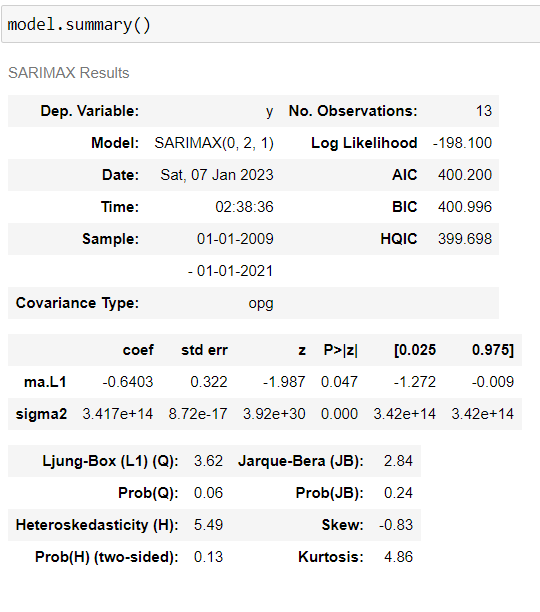


Figure 14 Suggested model

Prediction made using the arima model above generated some negative values. This suggests that the model is not good enough to use as a predictor. A sarimax model of order(0,1,1) was later fitted. This showed to be a promising model as the plot of the predictions and the real data look similar.

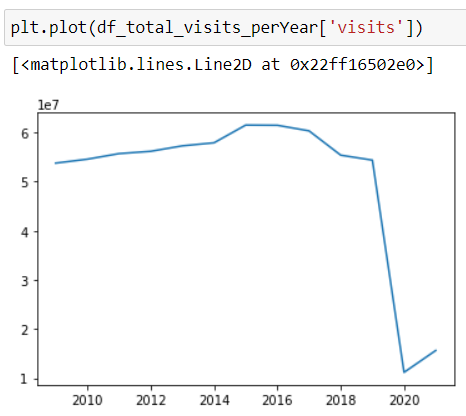


Fig 15:Plot of total visits per year

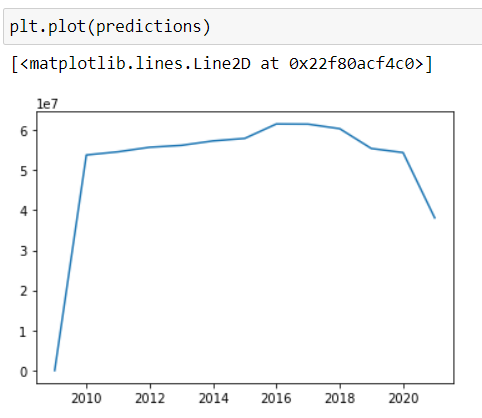


Fig 16:Plot of predicted values of total visits per year

**Forecasting**

Forecasting results for expenditure for the next 5 years (Year 2022 to 2027), it is observed that year 2023 will have the minimum expenditure amount of over 13 billion, anything less is not seen while 2027 is expecting over 16 billion

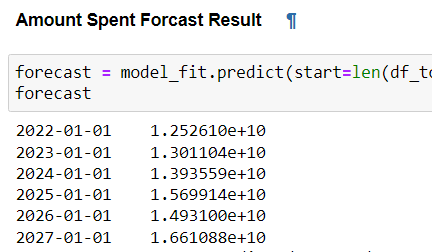


Figure 17: Expenditure Forecast Results

The same approach is being implemented in the total of visits per year, to forecast the number of visits for the next 5 years. There will be an increment after the year 2021. From the analysis being implemented, a static value was derived after the year 2020, which shows that the population of travelers would increase after the year 2022 and that would affect the total number of nights spent of the travelers to increase.

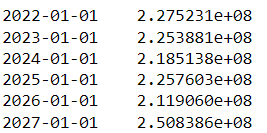


Figure 18: Forecast for Visits

Analyzing the number of nights spent using the SARIMAX model to forecast the number of night spent in the next 5 years. It shows that there is an increment after the year 2021. From the analysis being implemented, a static value was derived after the year 2020, which shows that the population of travelers would increase after the year 2022 , have ups and downs through the year 2023 to 2027, with 2027 estimate of over 250 million trevellers

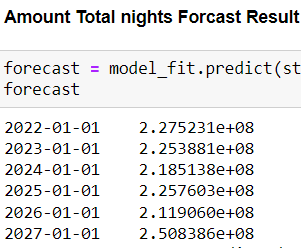


Figure 19: Forecast for Nights spent

# 4.4 Machine Learning Model Fitting

Machine learning is a method of teaching computers to learn and make decisions based on data, without explicitly programming them. It involves feeding a computer system a large amount of data, and allowing the system to use statistical analysis to identify patterns and make decisions. There are many different techniques and approaches to machine learning, including decision trees, support vector machines, and neural networks. These techniques can be used for a wide range of applications, such as image and speech recognition, natural language processing, and predictive modeling.

There are many different types of machine learning algorithms, but they can generally be grouped into one of four categories:

Supervised learning: This type of algorithm is trained on labeled data, where the correct output is provided for each example in the training set. The goal is to make predictions about unseen data based on the patterns that it learned from the training data. Examples of supervised learning algorithms include linear regression, support vector machines, and decision trees.

Unsupervised learning: This type of algorithm is not given any labeled training data and instead must find patterns and relationships in the data on its own. One common application of unsupervised learning is clustering, where the goal is to group similar examples together. Examples of unsupervised learning algorithms include k-means and hierarchical clustering.

Semi-supervised learning: This type of algorithm is trained on a dataset that is partially labeled and partially unlabeled. The goal is to use the labeled data to make predictions about the unlabeled data.

Reinforcement learning: This type of algorithm learns through trial and error, by taking actions in an environment and receiving rewards or penalties for those actions. The goal is to maximize the cumulative reward over time.

There are many variations and nuances within these categories, and different algorithms are better suited for different types of tasks and data. In this report, we will use the supervised machine learning to fit a model. The process of supervised machine learning generally involves the following steps:

Collect and preprocess the data: This includes gathering the relevant data, cleaning it to remove any errors or inconsistencies, and formatting it in a way that the algorithm can understand.

Split the data into a training set and a test set: The training set is used to train the model, while the test set is used to evaluate the model's performance.

Choose an appropriate model and training algorithm: There are many different models and algorithms to choose from, and the appropriate choice depends on the nature of the data and the task at hand.

Train the model: This involves feeding the training data to the model and using an optimization algorithm to adjust the model's parameters so that it can make predictions that are as accurate as possible.

Evaluate the model: The model's performance is evaluated using the test set. Common evaluation metrics include accuracy, precision, and recall.

Fine-tune the model: If the model's performance is not satisfactory, various techniques such as hyperparameter optimization can be used to improve the model.

Make predictions on new data: Once the model is trained and fine-tuned, it can be used to make predictions on new, unseen data.

For this work a supervised learning would be used.

**4.41 Division of data**

After thorough cleaning of the original dataset, the outliers were sieved out. The dataset is then divided into two subset, the Holiday package is the Y-data while the remaining data such as year, quarters, ages, number of nights, purpose are our X-data. Each of the X-data and Y-data is then split into Y-train, Y-test, X-train and X-test. The X data contains both categorical data and continuous data. The categorical data needs to be transformed into numeric format, dummies , that the computer can understand.

# 4.42 Model Fitting for Holiday Package

An initial accuracy test was for potential model was conducted using K-fold evaluation. Logistic regression and random forest classification have the highest initial accuracy ,as random forest have the highest accuracy, it was chosen. After fitting the a model with it, the model performed well with a model score of over 85% .

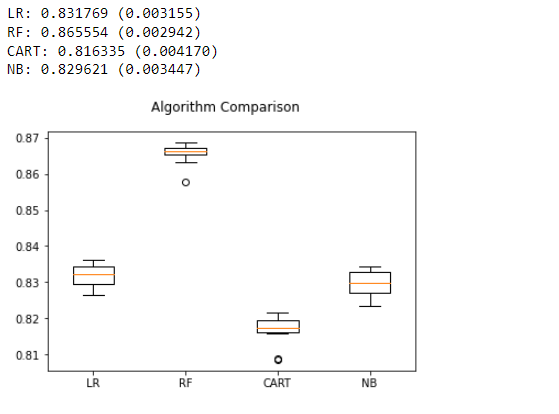


Fig 20: Initial accuracy test to choose a good model

4.43 Model accuracy of Holiday package.

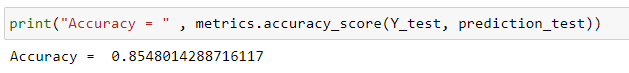


Fig 21: Model accuracy for Holiday Package

4.44 Feature importances of the Holiday package model

sklearn has a method called feature\_importances. Feature importances show the respective contribution of each of the feature and sub-feature to the accuracy of the model.

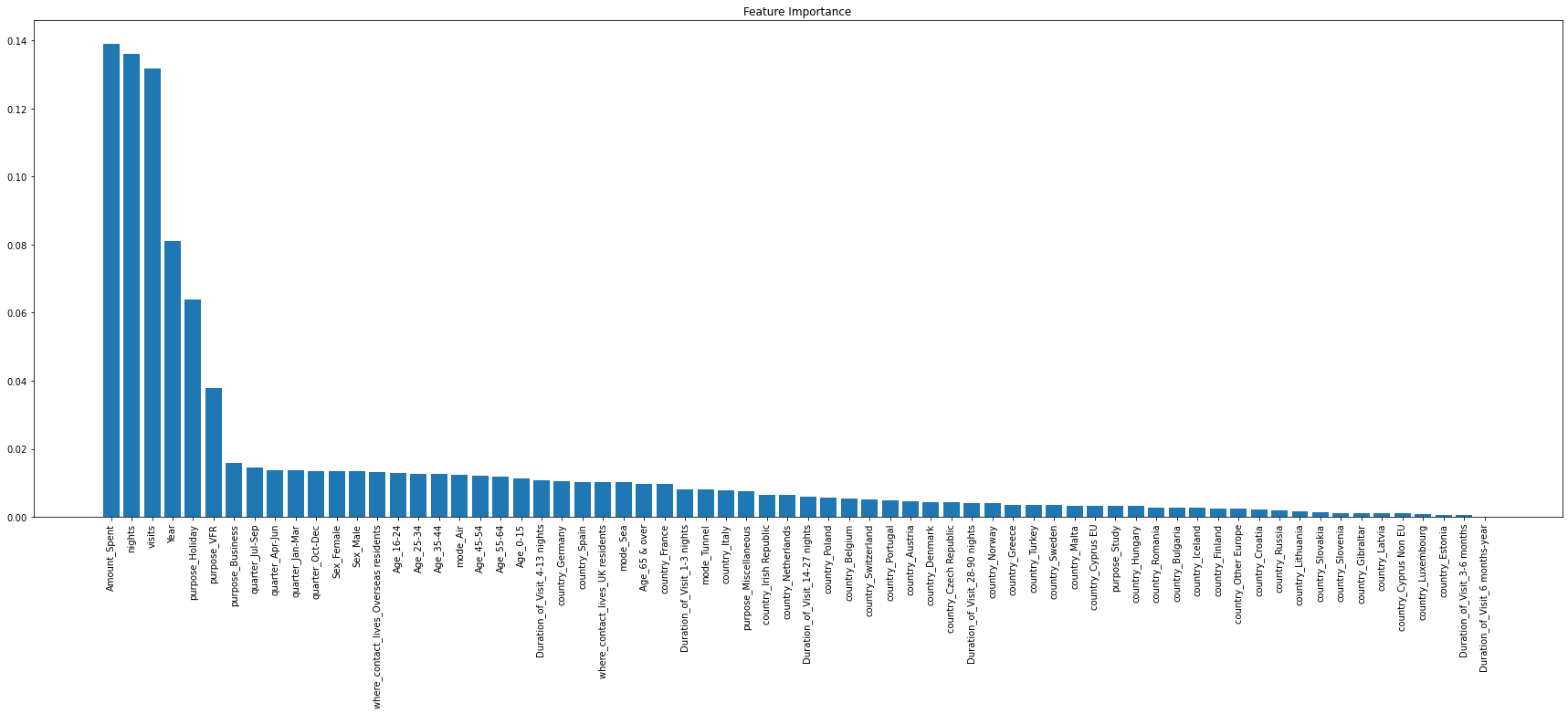


Fig 22: A barchart showing the features importances of each dependent variables to Holiday package in descending order

The best five features that contribute to the accuracy of the model are shown below.

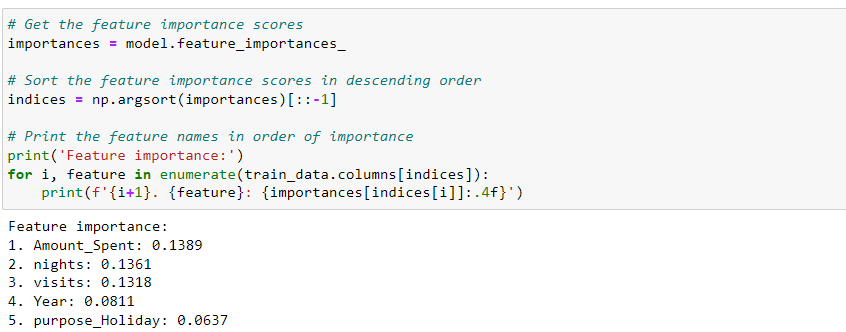


Fig 23: The best five features that contribute to the prediction of Holiday package

This showed that the amount spent contribute 13% to predicting Holiday package, nights 13%, visits 13%, year 8% and purpose\_Holiday, which is a sub-feature of holiday 6%, are the best predictor of the type of holiday package a traveler will have.

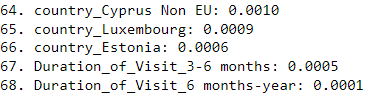


Fig 24: the least five features that contribute to prediction of Holiday package

4.45 Feature engineering

In an attempt to increase the accuracy of the machine learning model, two other models were fitted. A second model was fitted using all the features except duration of visit, the model have an accuracy of 85.1 % compared to the first mode with an accuracy of approximately 85.5% , 0.4% difference. This means that duration of visits(number of nights) is almost insignificant in predicting holiday package. The third model was fit omitting duration of visit and mode of transportation. It eventually have an accuracy of 84.1%, a 1.4% difference with the first model. Since both the second and third model didn’t increase the accuracy, we will stick to the first model.

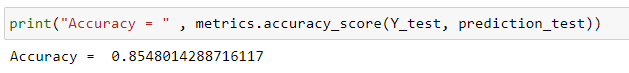


Fig 25: Model accuracy for first model using all features

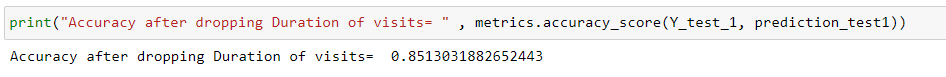


Fig 26: Model accuracy of the second model after dropping duration of visit

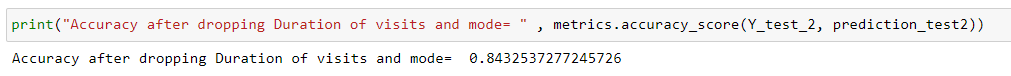


Fig 27: Model accuracy of the third model after dropping duration of visit and mode of transportation

4.5 Model fitting for Amount Spent(expend)

After fitting a model and making prediction with a time series ARIMA. We fitted a machine learning model considering the comparative advantages below.

Machine learning models and ARIMA (AutoRegressive Integrated Moving Average) models are both techniques used for time series forecasting. They both have their own strengths and limitations, and the appropriate one to use depends on the specific problem at hand.

One advantage of machine learning models is their ability to automatically learn complex patterns in data and make predictions based on those patterns. They can handle a large number of input features and can often make more accurate predictions than simpler models like ARIMA.

On the other hand, ARIMA models are simpler and easier to implement, especially for users who are not familiar with machine learning. They are also well-suited for forecasting when there is a clear trend or seasonality in the data.

In general, machine learning models may be a better choice when there is a large amount of data available and there is a need for high prediction accuracy. ARIMA models may be a good choice when the focus is on understanding the underlying patterns in the data rather than making the most accurate predictions possible.





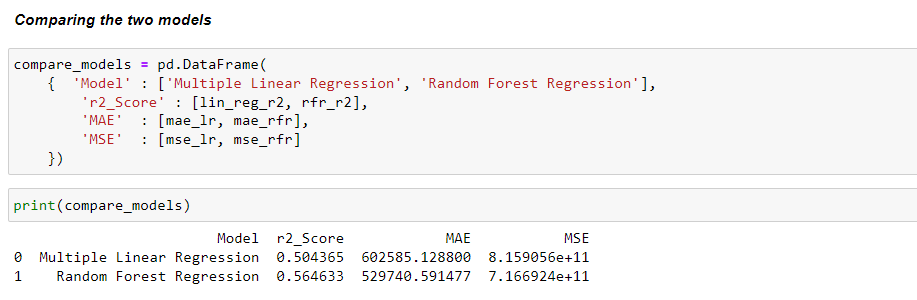


Fig 28: comparing model score

We can now see the score and error of Linear regression and Random forest classifier and compare them. Score of Random forest Regression is greater than Linear regression and error is also less. Thus, Random forest Regression will be the right choice for our model.

The r squared score of our machine learning model is 0.56. This means that only 56 percent variations in amount spent(expend) by travelers can be explained by all the features. This model is not good enough for prediction. In an attempt to make the model better , an effort was made to determine the relationship between some independent variables and amount spent as seen below. Pearson correlation coefficient was used for calculating the correlation coefficient between amount spent and other continuous variables like nights, visits and year.

While covariance was used to find extent of relationship between amount spent and other categorical variables like age and purpose as shown below.

4.51 Feature selection for Amount spent(expend)

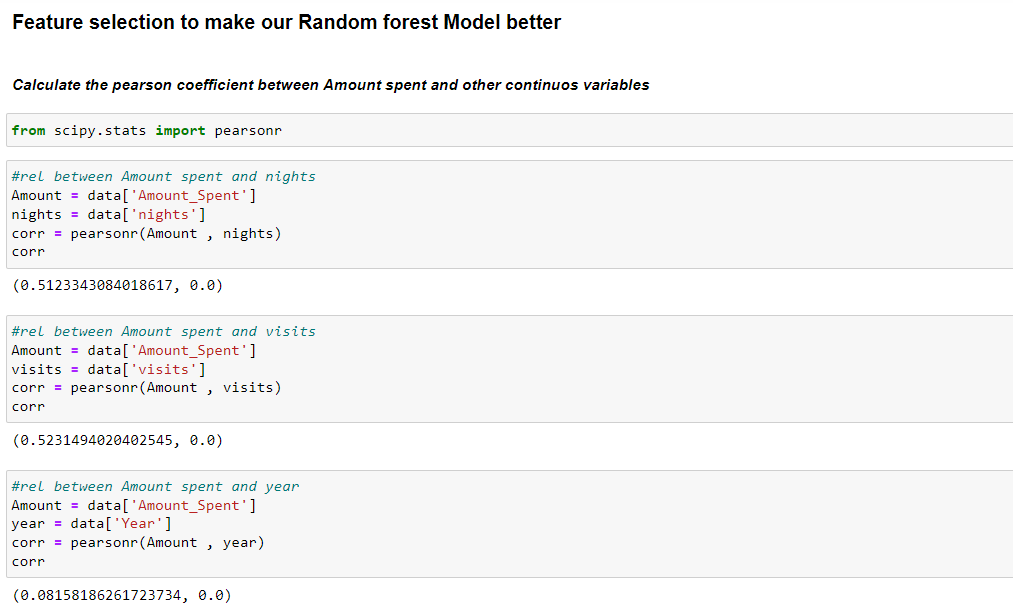


Fig 29: pearson correlation coefficient showing correlation between Amount spent and nights,visits and year

Result: Pearson correlation coefficient shows a high correlation between Amount spent and nights , visits and a weak correlation between amount and year.

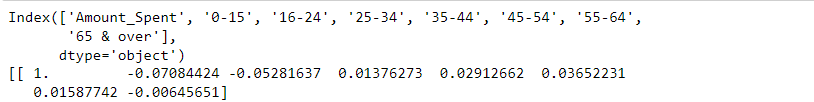


Fig 30: correlation matrix showing extent of relationship between Amount spent and Age

Result: The amount spent has a weak negative correlation with ages 0-15 , 16-24 and 65&over , While it has a weak positive correlation with the remaining ages, i.e ages 25-34 , 35-44 , 45-54 and 55-64. In other words , Amount spent is slightly increased if the individual is a working class , while it reduces with more non working class

Conclusion: Amount spent has a weak correlation with ages



Fig 31: Showing the covariance result between Amount spent(expend) and purpose of travelling.

Result: A weak positive relationship exist between Amount spent and Business , holiday and Study, while a weak negative relationship exist between amount spent and those who travel for visitation and Miscellaneous

Conclusion: An increase in the number of those who travel for Business, Holiday and Study translates to an increase in amount spent. An increase in the number of those who travel for visitation and Miscellaneous translates to a decrease in Amount spent

4.52 Comparing Model score after removing some features

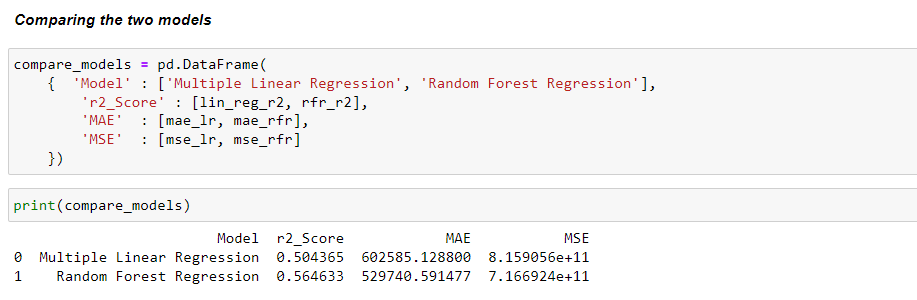


Fig 32: Model score of the first model before removing any feature

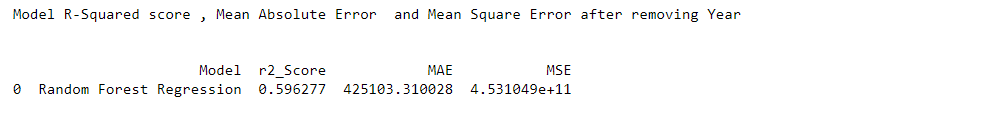


Fig 33: Model score of the second model after removing year

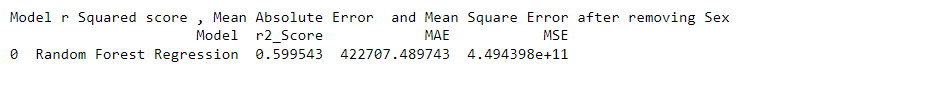


Fig 34: Model score of the third model after removing Sex

Conclusion: Since our model score is not becoming better, the best score so far is 60% accuracy. The result from its prediction cannot be accurate. For that, its better to stick with the prediction we got from our time series model.

# 4.4 Power Bi Dashboard Design

Power Bi is a business intelligence software that is used as a data visualization tool. It allows data to be imported from different sources and convert it to an efficient dashboard for proper data visualization. In this experiment, the travelpac dataset was imported and different visualization chart was pulled out from it.

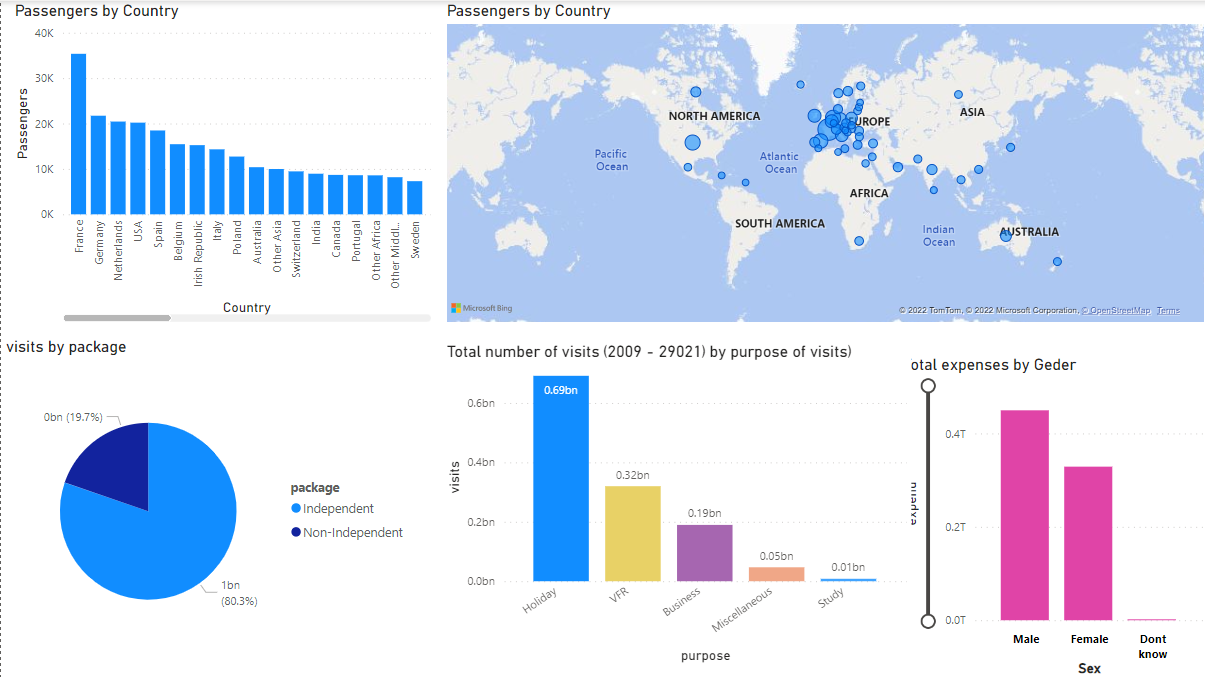


Figure 16: The Dashboard Interface

Figure 16 is the overall view of the entire interface, it consist of the stacked column, map and several data pie charts

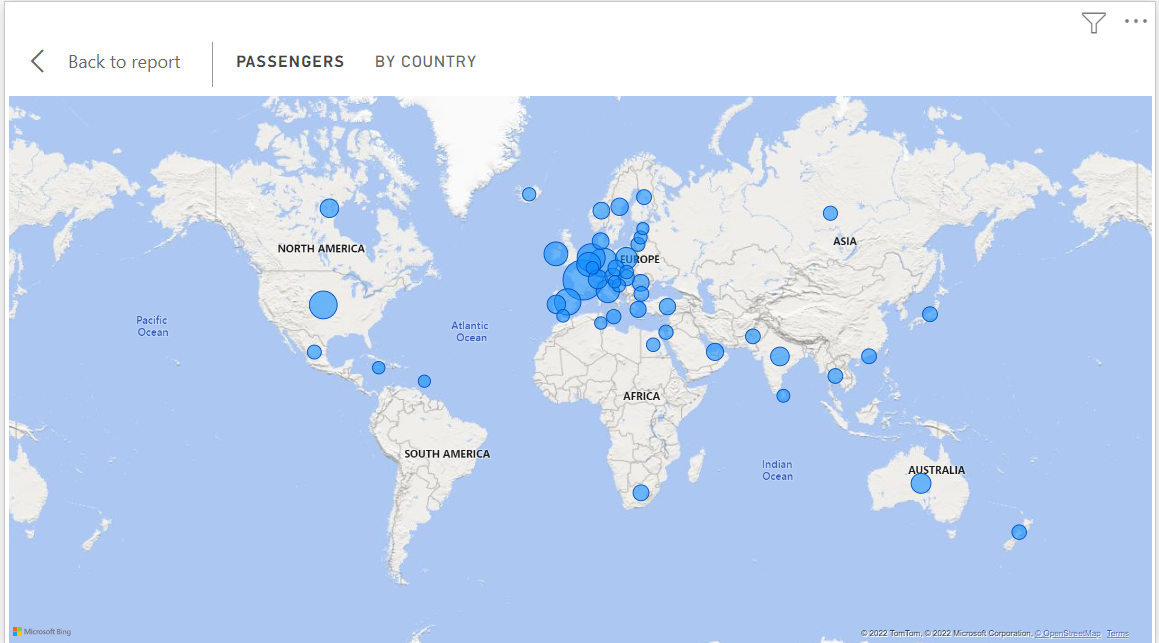


Figure 17: The Map

The Map is plotted against the country where the passengers are living. It is indicated with a blue circle as seen on the map.

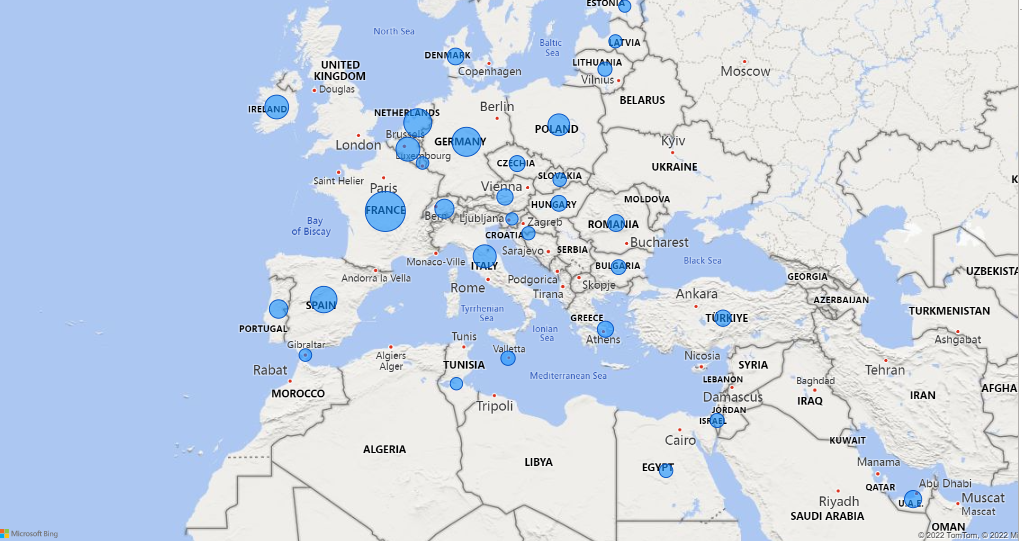


Figure 18 Zoomed Map

From figure 18 above, it shows that France has the highest number of travelers.

A pie chart showing the packages which the passengers used mostly, as seen clearly in figure 19, the independent package has the highest number of usage over non-independent

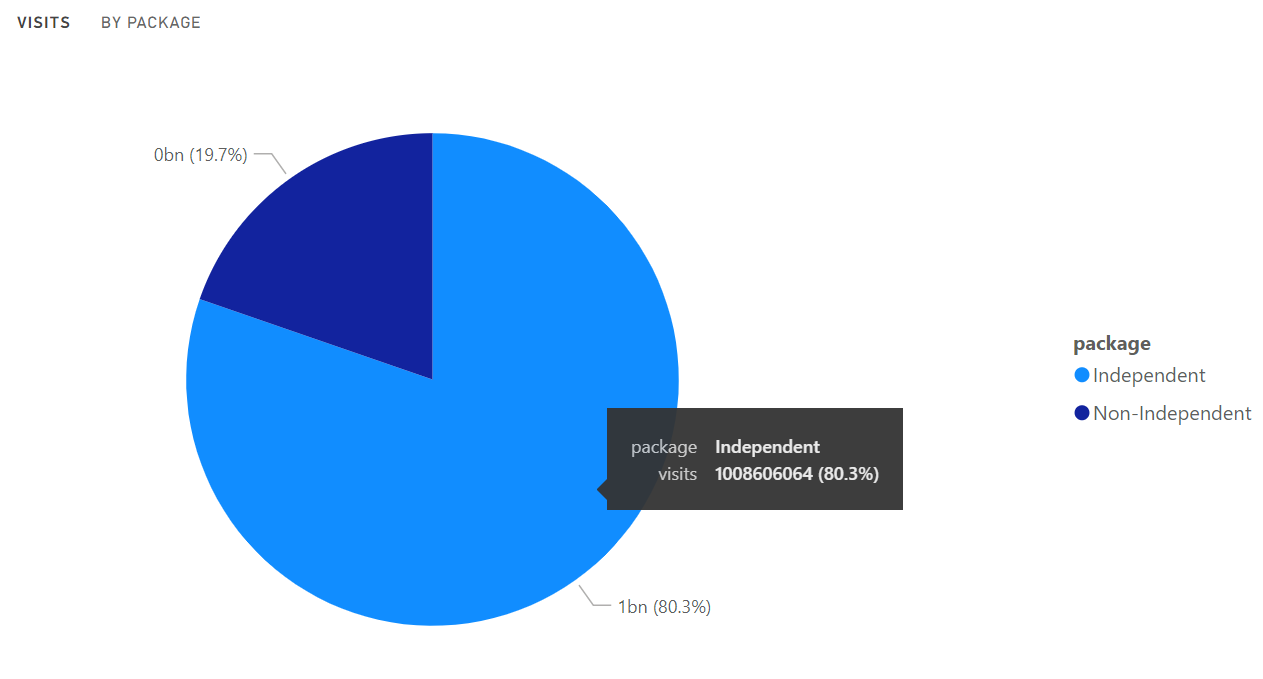


Figure 19: Packages Pie Chart

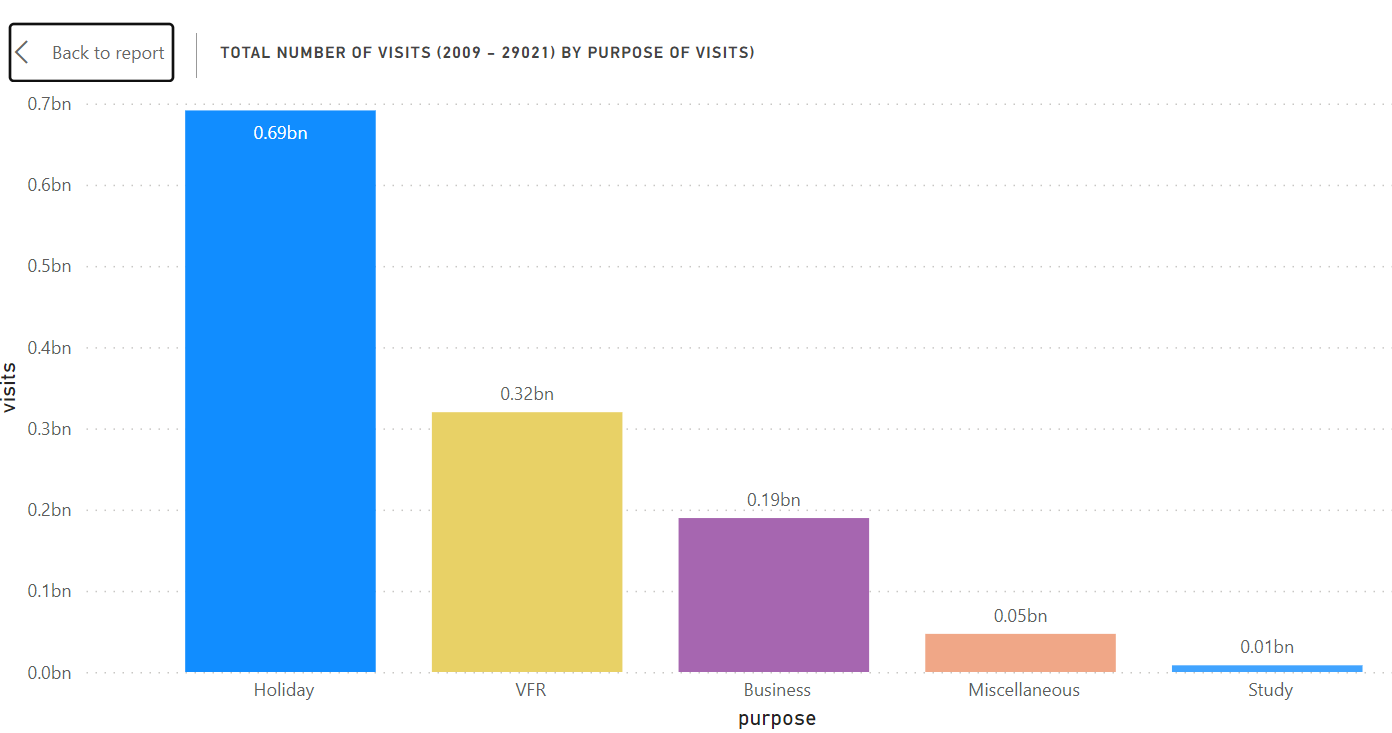


Figure 20: Travelling purpose

Another visualization used is the bar chart, showing the total number of visits and their purpose of visits, from the year 2009 to 2021. From the figure above, it shows that most of all the travelers from the year 2009 up to 2021 are travelling for holiday and the second most travelled purpose is the VFR (Visit Friend and Relatives), where passenger travel to visits either friend or their relatives. Travelling for business, miscellaneous and study purpose are the last respectively.

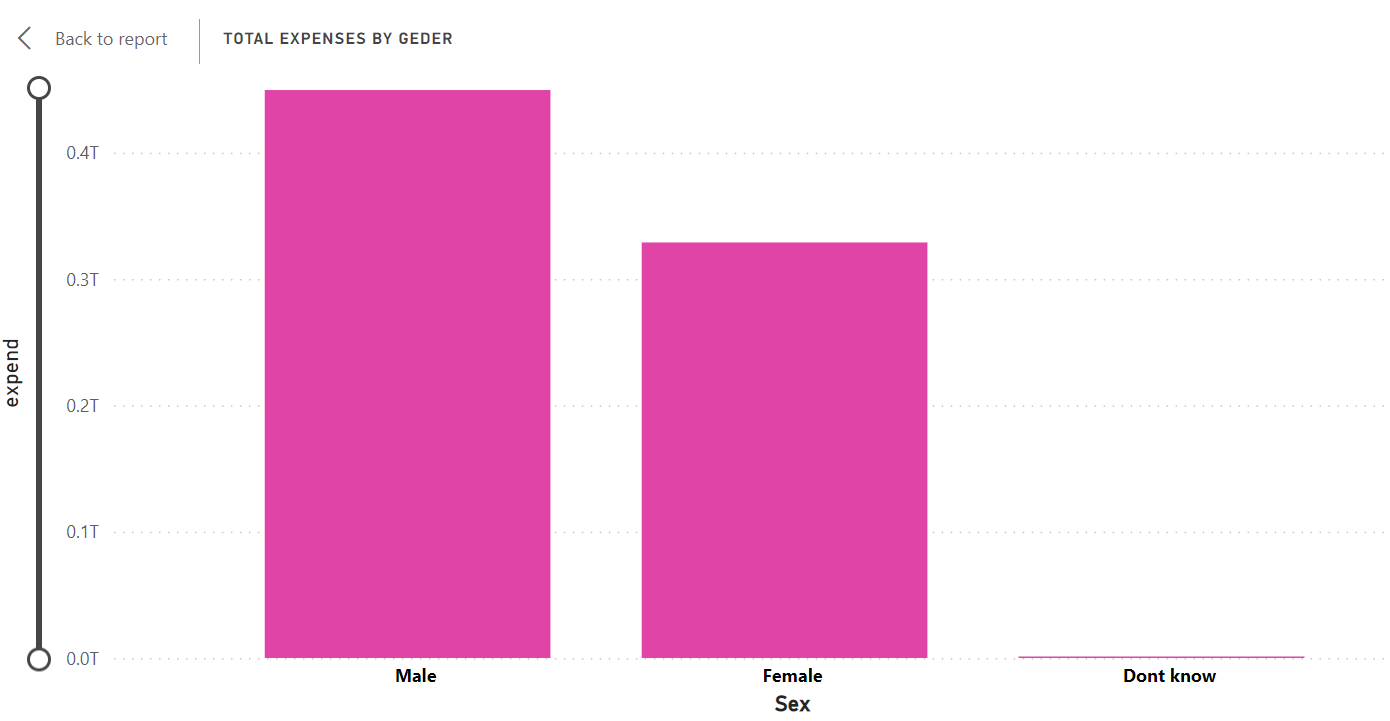


Figure 21: Gender Category

Figure 21 also shows we have more male traveler compare to the female travelers.

# 

# 5.1 Conclusions

Travelling into the UK can be easy for a passenger but as a travelling agency, it is their duty to ensure the transportation is not vulnerable to any attacks. With the survey and reviews conducted, its shows that there is a limited platform to gain access to information about travelling passengers in and out of the UK or perhaps let’s say it isn’t for the public to benefit.

Such information can be useful to researchers or to smaller travelling agencies that do not have the capacity to get the information themselves. More so, lack of information about passengers can make the travelling agencies vulnerable to attacks, having a better understanding about the passengers and the travelling can guide against any negative effects.

These challenges lead to the development of this study. To better solve the problems, we asked some research questions, such as.

i. What are the insights related to the travelers as regards their country, night spent, mode of travel and package of travel?

ii. What necessitates travelling?

iii. Does age groups affects travelling?

Two approaches were made. The first one involved using python programming libraries and analyses to fit time series model and machine learning models for some variables. Timeseries models forecasted the expected yearly outcomes of some of our variables for five years. Secondly,a dashboard was created using Power Bi using a dataset collected. Insights can then be viewed from the dashboard.

The summary of insights generated are as follows:

2010 has the highest number of data samples, while 2020 has the least

There are more oversea travelers than the UK travelers

Travelers who lives in the UK and overseas travelers mostly spent between 4 to 13 nights.

Further segmentaion of travellers who spent 4-13 nights revealed that for travelers that stay in the UK, the major reason for travelling is for holiday, the second major reason is visitation. Most overseas travellers are traveling for visitation purposes while the second major reason is for business.

To see the kinds of people who travel for visitation. The segmentation shows that travellers between the ages of 25 and 65 mostly go for visitation. This is very relatable as the age categories cover university graduates who might want to visit home when school are on break or adults who are travelling back home for visitation.

From age 25 – 55, we have larger number of passengers in that age group, considering that we have more passengers who travel for visitation and holidays, we can conclude that most passengers between the age of 25 – 55 are travelling for holidays and visitation purpose, more reason passengers are spending 1 to 13 nights during their visits

Categorizing the dataset into years and identifying the number of visits yearly, starting from 2009 till 2019, There is an increasing trend in the number of passengers but a huge drop in the year 2020, this is basically caused by COVID-19 lock-down and in the year 2021, it started increasing gradually

The place of residence for most overseas residents or of visit for most UK residents is France, with 35,021 passengers while Cyprus, which is on the least has 690 passengers.

The best four attributes that contribute to the prediction of the type Holiday package a traveler will have are: amount spent(expend), number of nights spent, number of visits and year.

Forecasting results for expenditure for the next 5 years (Year 2022 to 2027), it is observed that year 2023 will have the minimum expenditure amount of over 13 billion, while 2027 is expecting over 16 billion.

The model score machine learning model is 0.56. This means that only 56 percent variations in amount spent(expend) by travelers can be explained by all the features. This model is not good enough for prediction, this suggests that to predict the amount a certain traveler with some travel qualities correctly, we need more attributes.

There is strong relationship between Amount spent and nights , visits while we have weak relationship between amount and year.

The amount spent has a weak negative correlation with ages 0-15 , 16-24 and 65&over , Amount spent is slightly increased if the individual is a working class , while it reduces with more non working class.

An increase in the number of those who travel for Business, Holiday and Study translates to an increase in amount spent. An increase in the number of those who travel for visitation and Miscellaneous translates to a decrease in Amount spent

Several other insights were derived and were used in creating the dash board.

One key thing I learnt during the development of this project is consistency, initially my writing and programming skills was so amateur. But while progressing deeper into the report, I was able to revert back to what I have written before to make correction to what I have newly learnt about dissertation writing. Also, my python knowledge was minimal, I was able to go through series of programming tutorial to get some things done. Numerous code samples were reviewed, which helped in the analysis of this experiment. Also, for the development of the dashboard, this would be the first time I would be working with Power Bi, its an interesting tool to work with. This project introduced me to different tools and techniques as well.

Several problems were encountered, one of the big problems faced was the process of setting up Power Bi. The software is owned by Microsoft and can only be installed on a windows system. I am using a mac book air which requires I make a separate partition to install windows OS in other to install Power Bi. Another problem are the skills needed to complete the analysis. Several programming tutorials were downloaded and watched repeatedly to understand the concept of how python programming languages is being used and its application for data analysis.

During the development of the project, several mistakes were encountered such as programming error due to mistakes from wrong syntax. Also, in the dataset to be used. Initially, the dataset downloaded was just a year, at the point of completing the report, it shows that just a single year dataset can’t be used. All process was repeated over again, this time the dataset started from 2009 to 2021. This process reflected on the project, more insight was derived from this act which leads to yearly comparison of results.

In the future, this result in this analysis and the process can be converted into a web based application where information can be access by anyone. The dataset for the work is only for UK passengers, this can be extended to cover more regions.

# 

# References

Alit Suthanaya, P. (2018). Analysis of travel pattern and the need to develop sustainable transportation infrastructure in Sarbagita metropolitan area. *MATEC Web of Conferences*, *195*. https://doi.org/10.1051/matecconf/201819504017

Burns, E. (2021) *What is machine learning and why is it important?*, *Enterprise AI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML (Accessed: January 2, 2023).

*Covariance in Statistics: What is it? Example* (2021) *Statistics How To*. Available at: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/covariance/ (Accessed: January 2, 2023).

Drren Stilwell, J. C. M. S. (2018). *Statistical Release Analyses from the National Travel Survey*.

Ebner, J. (2020) “How to use Pandas reset index,” *Sharp Sight*. Sharp Sight, Inc, 11 February. Available at: https://www.sharpsightlabs.com/blog/pandas-reset-index/ (Accessed: January 2, 2023).

*Feature importances with a forest of trees* (no date) *scikit-learn*. Available at: https://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_importances.html (Accessed: January 2, 2023).

Halden, D. (2019). *UK Travel Time, Accessibility and Connectivity Statistics-Position Paper*. https://doi.org/10.13140/RG.2.2.13077.76005

Harms, T., Gershuny, J., & Olaru, D. (2018). Using time-use data to analyse travel behaviour: Findings from the UK. *Transportation Research Procedia*, *32*, 634–648. https://doi.org/10.1016/j.trpro.2018.10.007

Pathak, P. (2020) *How to create an ARIMA model for time series forecasting in python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/ (Accessed: January 2, 2023).

(No date a) *Openai.com*. Available at: https://chat.openai.com/chat (Accessed: January 2, 2023).

(No date b) *Researchgate.net*. Available at: https://www.researchgate.net/post/What\_is\_the\_minimum\_value\_of\_correlation\_coefficient\_to\_prove\_the\_existence\_of\_the\_accepted\_relationship\_between\_scores\_of\_two\_of\_more\_tests (Accessed: January 2, 2023).

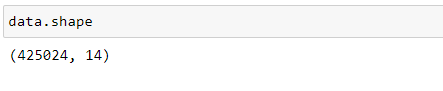
Yuniawati, Y., & Ridwanudin, O. (2015). Analysis of Travel Experience Quality at City Destinations. *Journal of Business on Hospitality and Tourism*, *1*(1), 8. <https://doi.org/10.22334/jbhost.v1i1.15>

# Appendix

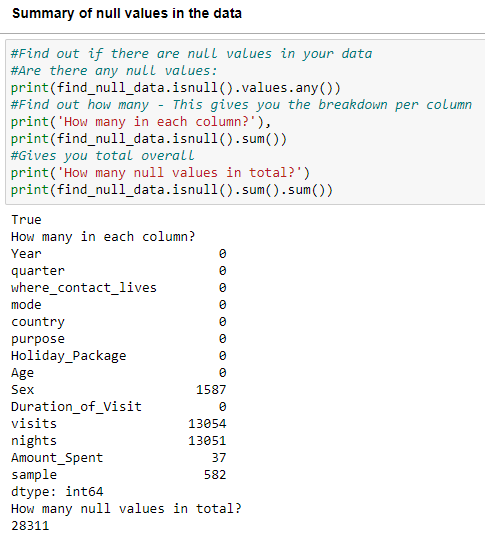
Raw dataset before cleaning, this dataset contains cells with empty values.



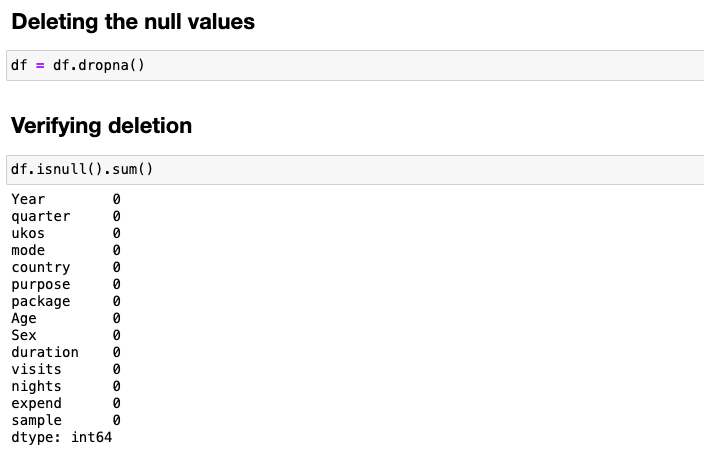
The figure below shows the shape of the data before data cleaning



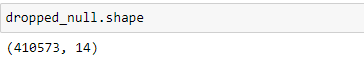
The figure below shows that column sex, visits, nights and sample contain error values in its columns. This shot was taken before the dataset was cleaned



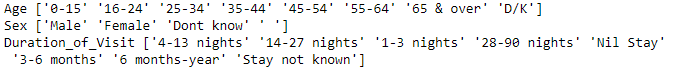
After the error values have been identified. It has to be removed.



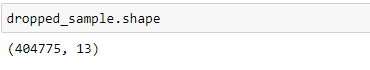
Shape of the data after deleting null values



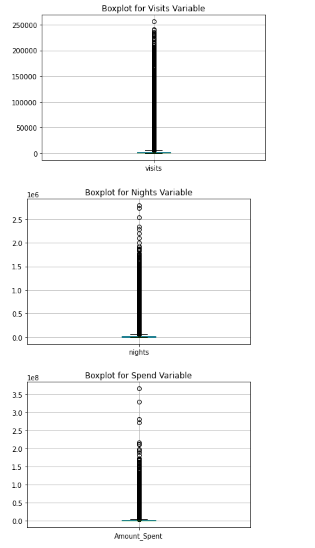
The figure below shows the presence of some irrelevant data.



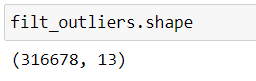
Shape of the data after data cleaning.



Box plot showing presence of outliers in the cleaned data

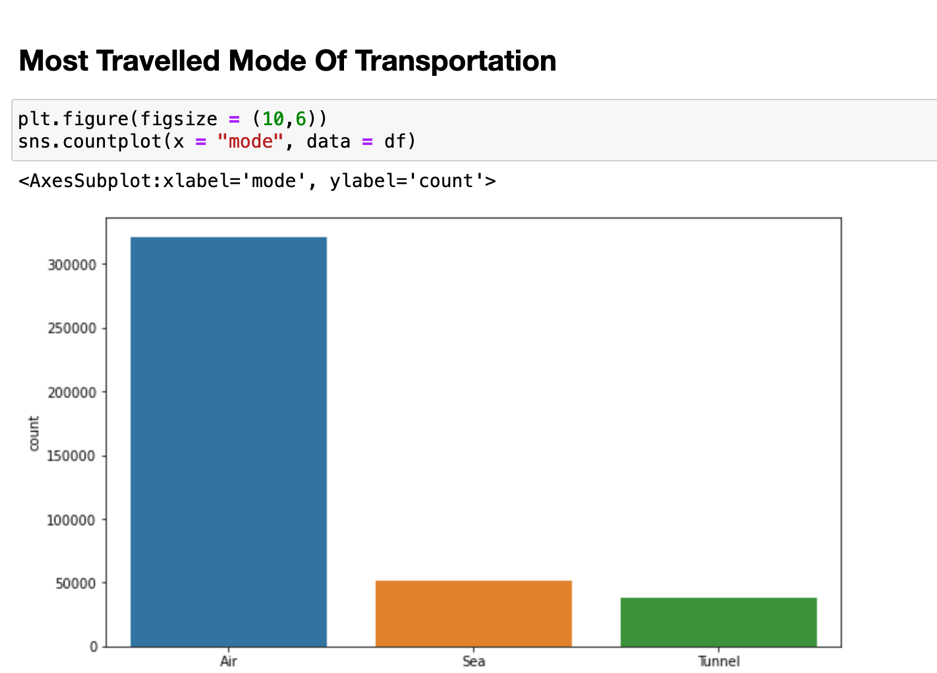


Shape of the data after outliers have been removed

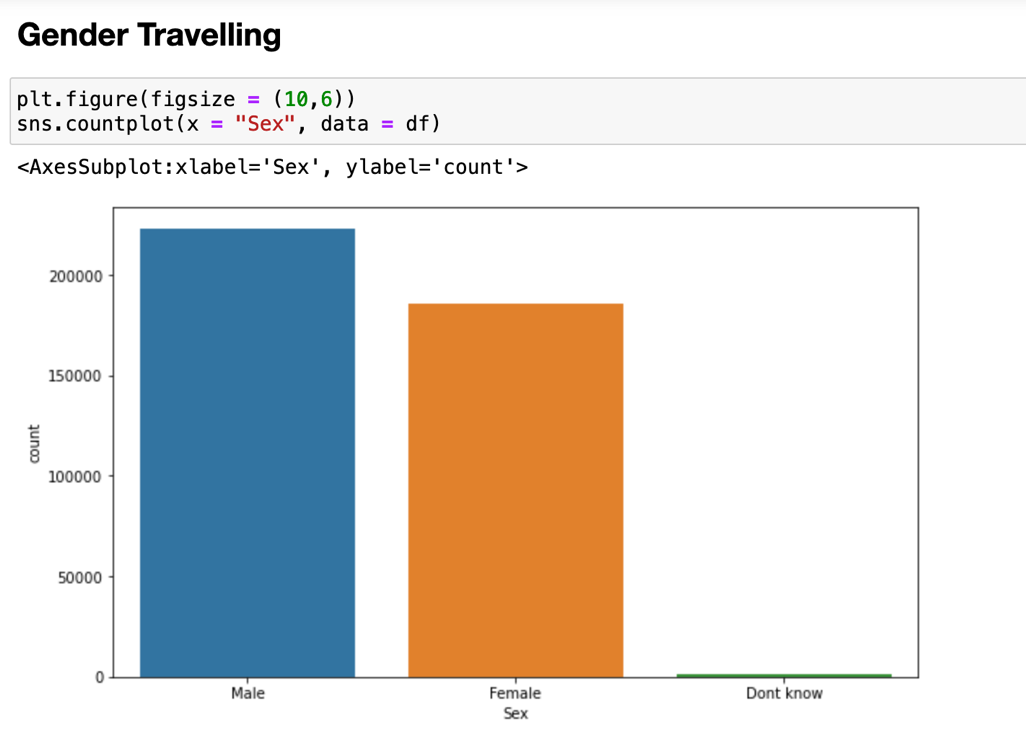


**Insight Snippets**

1. The most traveled mode of transportation is Air.



1. Most Travelled Gender



1. Most Used Package for Travelling

